Cancer Epidemiology 50 (2017) 207–213

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Detection and estimation of the increasing trend of cancer incidence in relatively small populations

Rina Chen[a,*], Enrique Y. Bitchatchi[b]

[a] BioForum, Applied Knowledge Center, Ness-Ziona, Israel
[b] University of Gerona, Gerona, Spain

ARTICLE INFO

Article history:
Received 27 October 2016
Received in revised form 2 April 2017
Accepted 6 April 2017

Keywords:
Joinpoint regression
Trend in cancer incidence
Temporal clustering
Cuscore test
Relative interval (RI)
q-interval

ABSTRACT

Background: Detection and estimation of trends in cancer incidence rates are commonly achieved by fitting standardized rates to a joinpoint log-linear regression. The efficiency of this approach is inadequate when applied to a relatively low levels of incidence. We compared that approach with the Cuscore test with respect to detecting a log-linear increasing trend of chronic myelomonocytic leukemia (CMML) in datasets simulated to match a province of about 700,000 inhabitants.
Methods: For better efficiency, we replaced the standardized rate as the dependent variable with a continuous statistic that reflects the inverse of the standardized incidence ratio (SIR). Both procedures were applied to datasets simulated to match published results in the Girona Province of Spain. We also present the use of the *q-interval* in displaying the temporal pattern of the events. This approach is demonstrated by analyses of CMML diagnoses in Girona County (1994–2008).
Results: The Cuscore was clearly more efficient than regression in detecting the simulated trend. The relative efficiency of the Cuscore is likely to be maintained in even higher levels of incidence. The use of graphical displays in providing clues regarding interpretation of the results is demonstrated.
Conclusions: The Cuscore test coupled with visual inspection of the temporal pattern of the events seems to be more efficient than regression analysis in detecting and interpreting data suspected to be at elevated risk. A confirmatory analysis is expected to weed out 75% of the superfluous significant results.

1. Introduction

Prediction of cancer incidence in a given population is often based on fitting a joinpoint regression model [1–4] to the logged standardized rates. This model assumes several consecutive linear trends (connected at joined points). The predicted incidence is based on the estimated annual percentage change (APC) derived from the slope of the last part of the joinpoint regression. Although it is well known that the efficiency of this procedure is inadequate when applied to small numbers of incident cases [3,5], it has been applied in such situations even when no incident cases were observed [4,5].

That inefficiency may be related to the difficulty in complying with linearity and with other restricted conditions underlying regression. In contrast, the efficiency of temporal clustering techniques is based only on appearance of clustering among some consecutive cases. Such clustering is expected under any form of increased rate (linear or not). Based on that reasoning, we compared the efficiency of the Cuscore test [6–9] with that of regression in detecting a log-linear increasing rate in data of small incident numbers. The datasets were simulated to match published results of chronic myelomonocytic leukemia (CMML; ICD-O-3 code: 9945/7) in Girona province, Spain [4].

We replaced the usual dependent variable in the regression analyses with the *RI* (relative interval) statistic. *RI* measures the time intervening between two consecutive events where each event includes a predefined number (r) of consecutive cases. Being a continuous positive variable, it is somewhat more stable than any adjusted rate (as the random variable of any adjusted rate is a function of the annual observed number of cases) and bypasses the need to deal with zero observed cases in some years.

* Correspondence to: 38/33 Hanassi St., Tel. Aviv 69206, Israel.
E-mail addresses: rinachen@netvision.net.il (R. Chen), eboccupenviron@gmail.com (E.Y. Bitchatchi).

http://dx.doi.org/10.1016/j.canep.2017.04.005
1877-7821/

2. Data

2.1. Simulated datasets

Simulation of 300 datasets was carried out assuming 61 cases diagnosed in a 15-year stable population with respect to size and risk factors. Under the geometric series with an annual increase of 3.3%, the number of cases in the first year was calculated to be $n_1 = 3.21$.

The simulation was carried out for 60 cases, assuming that the expected number of new cases in year t is: $E(X_t) = 3.21^{*}1.033^{t-1}$. Accordingly, the expected interval between consecutive cases is $E(W_t) = 12/E(X_t)$ months. Assuming exponential distribution, the number of months between consecutive diagnoses was randomly allocated (using STATA [10]). When an interval extended over 2 years, we used a rule (described below) whereby the expected interval was determined as either $E(W_t)$ or $E(W_{t+1})$. It should be noted that either choice leads to a biased allocated time interval. However, for data covering 15 years, the bias will be in intervals of 14 (out of 60) cases at most, and will affect the waiting time of only 1/r cases of the relevant event. In order to minimize bias, the choice between $E(W_t)$ and $E(W_{t+1})$ was made such that most of the interval is allocated under the correct $E(W_t)$. By our rule of thumb, the allocated time interval was according to $E(W_{t+1})$ if the time remaining in year t after the diagnosis date of the previous case was less than $E(W_t)/4$.

The analyses are based on the recorded time of events where each event includes r consecutive cases. The three r values are: 3, 4 and 5. We grouped the data into three r groups (each with 100 datasets) according to the size of r.

2.2. CMML cases in Girona County (1994–2008)

Our approach is demonstrated using real data recorded during a 15-year period in Girona County.

2.2.1. Girona County Region (the Central Comarca of the Province of Girona)

Girona County constitutes about one-fourth of Girona province's population and half of its latitude. The community of the county is better off than the community of Catalonia *en bloc* with respect to economic welfare and healthcare availability. In general, residential communities are quite stable (in size and profile) over the 15-year period. However, an influx of young people of working age began in the mid-1990s. The possible effect of that immigration on the age profile of our analyzed data was found to be negligibly small [11,12].

2.2.2. Reference population

The best affordable a priori reference population was extracted as an aggregate of the annual counts over 36 age and gender strata and 221 municipalities of Girona Province. Catalonian Statistics Institute (IDESCAT) strata counts exist only for the last 10 years (1999–2008). However, municipal census could be accessed directly. The ultimate dataset originated from municipalities' census apiece plus smoothing splines generalized linear model (GLM) including Poisson response. This yielded a 15-year aggregated reference for 1994–2008.

2.2.3. Annual expected number of cases

Specific age and gender incidence rates were assumed to be the rates observed in the Girona Province during the 15-year period from 1994 to 2008. Denoting by R_S the age- and gender-specific rate in the reference population, and by $N_{S,t}$ the relevant group size in Girona County in year t, the expected number of new cases in year t is: $E(X_t) = \sum R_s{}^{*} N_{S,t}$.

3. Methods

3.1. Test statistics

3.1.1. The RI

The *RI* statistic is defined as w/E*(W)*, where w is the observed number of months intervening between two consecutive events and E(*W*) is the expected number of months between two consecutive diagnoses. Accordingly, the length of *RI* is simply the expected number of cases during w months. It can easily be updated with respect to temporal changes in the population's profile by updating annually the expected number of cases. Thus, *RI* is the waiting time until the event, measured by the expected number of cases regardless of the current E(W) length. As such its distribution is gamma and its mean is the event size r. The fact that the mean time until the event equals the event size is clear. This is so since the expected time until a single case is E(W) months, hence the expected time until r cases is r*E(W), thus RI = r*E(W)/E(W) = r. Practical details of RI calculation are demonstrated in 4.2.

It is interesting to note that when the annual expected number of cases is r, RI/r is the inverse of SIR, since RI/r = exp/obs ≈ 1/SIR. The difference between the two measures is the random variable. It is the observed number in SIR and the expected number of cases in *RI*.

Based on this, and in order to comply with the common practice in which trend analysis is based on the annual incidence, we suggest that r is defined as the upper integer of the annual expected number of cases at baseline. It is quite likely – even under an increasing trend – that the number of cases expected annually is still close to r. In our simulated data, analyses are based on r = 4. We also analyzed data in which the event included three or five cases. Results of these two r values provided better insight regarding the relative efficiency of the two tests.

The easy accommodation of *RI* under changing conditions, and the fact that its gamma distribution depends only on r, enabled the derivation of several procedures aimed at detection and interpretation of the increased rate of cancer diagnoses [6–9,13–18].

3.1.2. The q-interval

The *q-interval* [17] is defined as the a-priori probability that the waiting time until the event is longer than that observed. Namely, it is the a-priori probability that the r^{th} case of an event is diagnosed after the observed *RI*. As a gamma distributed variable, the *q-interval* can also be calculated under the Poisson distribution. Under the Poisson it is defined as the probability that no more than r-1 cases are observed during an interval in which *RI* cases should be expected. For example, suppose that r = 4, E(W) = 3 months and w = 16 months. Namely, four cases were observed during a period in which (RI = 16/3 =) 5.33 cases should be expected. According to the Poisson distribution, the probability that no more than three cases are observed during *RI* = 5.33 is the *q-interval* = 0.222.

Although the *q-interval* is calculated as a probability value, it is actually a random variable derived from an observed random event (*RI*). It is a cumulative probability of a continuous random variable; as such, its distribution is uniform over 0–1 [19]. Accordingly, under stable conditions, its expected value is 0.5 (for any r value) and >0.5 under elevated incidence. Based on that distribution we can use the *q-interval* in graphical display of the temporal pattern of the events.

3.2. Analyses

Both regression and Cuscore procedures were applied to each of the simulated datasets. The relative efficiency of the two procedures was tested for significance (two-tailed) by applying McNemar's test [20].